AD-A011 709

STUDY AND DEVELOPMENT OF THE INTEL TECHNIQUE FOR IMPROVING SPEECH INTELLIGIBILITY

M. R. Weiss, et al

Nicolet Scientific Corporation

Prepared for:

Rome Air Development Center

April 1975

RADC-TR-75-108
Final Technical Report
April 1975

STUDY AND DEVELOPMENT OF THE INTEL TECHNIQUE FOR IMPROVING SPEECH INTELLIGIBILITY

Nicolet Scientific Corporation

Approved for public release;
distribution unlimited.

Laboratory Directors' Fund No. 01737401

Rome Air Development Center
Air Force Systems Command
Griffiss Air Force Base, New York 13441

Reproduced by
NATIONAL TECHNICAL
INFORMATION SERVICE
U S Department of Commerce
Springfield VA 22151

This report has been reviewed by the RADC Information Office (OI) and is releasable to the National Technical Information Service (NTIS). At NTIS it will be releasable to the general public, including foreign nations.

This report has been reviewed and approved for publication.

APPROVED: Robert A. Curtis

ROBERT A. CURTIS, Capt, USAF
Project Engineer

APPROVED: H Davis

HOWARD DAVIS
Technical Director
Intel & Recon Division

FOR THE COMMANDER: John P Huss

JOHN P. HUSS
Acting Chief, Plans Office

This effort was funded totally by the Laboratory Directors' Fund.

Do not return this copy; retain or destroy.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER RADC-TR-75-108	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) STUDY AND DEVELOPMENT OF THE INTEL TECHNIQUE FOR IMPROVING SPEECH INTELLIGIBILITY		5. TYPE OF REPORT & PERIOD COVERED Final Technical Report Sept 73 - Dec 74 6. PERFORMING ORG. REPORT NUMBER N/A
7. AUTHOR(s) M. R. Weiss E. Aschkenasy T. W. Parsons		8. CONTRACT OR GRANT NUMBER(s) F30602-74-C-0058
9. PERFORMING ORGANIZATION NAME AND ADDRESS Nicolet Scientific Corporation 245 Livingston Street Northvale NJ 07647		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 61101F 01737401
11. CONTROLLING OFFICE NAME AND ADDRESS Rome Air Development Center (IRAP) Griffiss AFB NY 13441		12. REPORT DATE April 1975 13. NUMBER OF PAGES 44
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office) Same		15. SECURITY CLASS. (of this report) UNCLASSIFIED 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE N/A

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

Same

18. SUPPLEMENTARY NOTES

RADC Project Engineer:
Robert A. Curtis, Capt, USAF (IRAP)
AC 315 330-2354
This effort was funded totally by the Laboratory Directors' Fund.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Speech enhancement
Speech intelligibility

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

A process, called INTEL, has been developed for improving the signal-to-noise ratio of speech obscured by wideband noise. The process reduces the level of the noise substantially without significantly attenuating or distorting the imbedded speech. At its present stage of development, INTEL's ability to enhance speech intelligibility is uncertain. However, it is effective in reducing listener fatigue, particularly when the S/N is below 6 dB. The process bears a superficial similarity to the homomorphic filtering techniques

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

of Oppenheim et al but differs from them in that it does not use the concept of generalized linearity.

It does not appear that this process has been pushed to the limit of its potential. Hence, in the present research, a theoretical analysis of the process has been made, and a number of experiments have been carried out, in an attempt to improve the performance of the process. The analysis yielded a statistical description of the response of INTEL to white noise and to vocalic speech, and a qualitative description of the response to speech and additive noise. The experiments centered about various modifications to the process which seemed likely to render the recovered speech more intelligible and more natural-sounding.

The experiments did not lead to distinct improvements in the process. However, the description provided by the analysis pointed the way toward other areas, not covered by the experiments, which look highly promising. In particular, improved methods of removing the noise term in the transformed signal may be successful. It is recommended that these lines be pursued in further research.

TABLE OF CONTENTS

LIST OF FIGURES

EVALUATION

This report describes theoretical and experimental studies done to improve the performance of INTEL, a process for enhancing the signal-to-noise ratio of speech which has been corrupted by wideband noise. The theoretical studies showed some of the statistical properties of the Gaussian noise in the INTEL process. While the intelligibility of speech is not mathematically definable, several experiments indicate that the rooting method used in part of the INTEL process tends to separate the speech and noise components.

While the phase is left alone in this process, experiments indicate that the correct phase will improve the intelligibility considerably. Future work should investigate possible methods of reducing the phase noise.

The modified cepstrum gating process, as a result of the theoretical and experimental work, should also be modified in future research.

The theoretical and experimental work performed indicated several promising areas for future research.

ROBERT A. CURTIS
Capt, USAF
Project Engineer

1.0 INTRODUCTION

This report describes theoretical and experimental studies done to improve the performance of INTEL, a process for improving the signal-to-noise ratio of speech which has been corrupted by wideband noise.

The theoretical study is a detailed statistical study of the process, showing how it works, why its effect on speech is different from its effect on noise, and why the cepstrum, a closely similar process, does not provide similar benefits. In addition, possible areas for further research are identified from the theoretical findings.

The experimental studies were explorations suggested by previous studies and do not reflect the findings of the theoretical investigation mentioned above. These studies implemented various modifications to the basic process in the hope that they would improve its performance. The modifications were:

Threshold clipping
Center clipping
Harmonic emphasis
Adaptation to narrow-band speech

In addition, a number of experiments were conducted to study the effect of phase on speech intelligibility.

2.0 ANALYSIS OF THE INTEL PROCESS

The noise reduction process known by the name of INTEL was largely developed empirically, as an attempt to capitalize on the difference between the autocorrelation functions of noise and voiced speech. (The motivation and early phases of the research are described in Ref. 1.) In order to continue improvement of the process, however, a theoretical understanding of the process is essential. Accordingly, we have done such an analysis, the results of which are given in this section. After showing how the process can be approximated by a simplified model, we analyze the behavior of the system with noise input and with speech input. We use the results of these analyses to provide a qualitative description of what happens when noisy speech is processed. (Because of the complexity of the system, we have not been able to provide a quantitative description in the case of noisy speech.)

2.1 Description of Process

The basic process is shown in block-diagram form in Figure 1. The process consists of the following steps:

1. The incoming signal is divided into overlapping segments 51.2 msec long. (Each segment is processed separately and the output segments are overlapped and added, producing the

processed signal, as described in Ref. 1. The steps given here detail the processing of one segment.)

2. Triangular weighting is applied to the segment. This reduces spectrum sidelobes during the process and, in the output speech, smooths the transition from one processing regime to the next.

3. The array is Fourier transformed.

4. The absolute value of the transform is taken. The phase is saved for future reference.

5. The upper half of the array is set to zero and the n^th root of the remaining part is taken. (n is usually 2 or 3.) We term this process <u>root compression</u>; n is called the <u>root compression factor</u>.

6. Every odd-numbered element is reversed in sign. (This is simply a computational convenience which has the effect of shifting the origin of the next transform to a more convenient location.)

7. The array is Fourier transformed a second time. The result of this second transformation is called the pseudo-cepstrum. (A regular cepstrum is generated by logarithmic compression instead of root compression.)

8. Samples adjacent to the origin are set to zero.

This operation, which is the essential step in the process, is called <u>gating</u>.

From this point on, the transformations of Steps 4 through 7 are undone: that is, the function is inverse-transformed, the sign reversals are removed, the array is raised to the nth power, the phase is restored, and a second inverse transform is done.

The purpose of the gating operation, Step 8, is to remove a large buildup around the origin which is due to the noise. When this is done, we find that the noise level in the reconstructed signal is significantly reduced. Clearly the sequence of steps from 4 through 7 can be regarded (except for the zeroing-out of the upper half of y_2) as a single reversible transformation. Accordingly, the first question to be asked is: why does this transformation (apparently) move most of the noise down to within a few samples of the origin and not do the same thing to speech? To answer this question, we first make a simplified model of the process and then analyze the behavior of the model when noise or speech signals are applied.

2.2 <u>Analysis with Noise Input</u>

The simplified model used in this analysis is formed by removing the time-weighting function (Step 2), the zeroing-out of the upper half of y_2 (Step 5), and the sign-reversal of alternate elements (Step 6). The removal of Step 6 has no significant effect of the process; the removal of time weighting and the zeroing-out operations make a small difference that can be easily corrected once we understand the process as a whole. A block diagram of the simplified model is given in Figure 2. From this figure, it will be seen that we use x to designate the input signal, y_1 to designate the first transform, y_2 to designate the absolute value of y_1, y_3 to designate the n^{th} root of y_2, and z to designate the second transform. When these signals are stochastic processes, this fact will be indicated by the use of a wavy underscore: e.g., $\underset{\sim}{y}_1$. For convenience, the independent variable will be omitted when no ambiguity will result from doing so. We will now describe the statistics of each of these signals when white noise is applied to the system. In this analysis, we will assume a familiarity with stochastic processes and with the properties of the discrete Fourier transform (DFT).

a. Input Signal. Let $\underset{\sim}{x}(t)$ be a sequence of real, zero-mean stationary Gaussian noise samples of amplitude $\sigma_{\underset{\sim}{x}}$. Then x has the probability density function,

$$f_{\underset{\sim}{x}}(x) = \frac{1}{\sigma_{\underset{\sim}{x}}\sqrt{2\pi}} \exp(-x^2/2\sigma_{\underset{\sim}{x}}^2) \qquad (1)$$

Since $\underset{\sim}{x}(t)$ is assumed white, its autocorrelation function is given by

$$R_{\underset{\sim}{x}}(\tau) = \sigma_{\underset{\sim}{x}}^{2}\delta(\tau) \qquad (2)$$

b. <u>First Transform</u> $(\underset{\sim}{y}_1)$. Following Ref. 2, p. 368f, we derive the statistics of $\underset{\sim}{y}_1(f)$, the DFT of $\underset{\sim}{x}$. By definition,

$$\underset{\sim}{y}_1(f) = \frac{1}{N}\sum_{t=0}^{N-1} \underset{\sim}{x}(t)\, W^{-ft} \qquad (3)$$

where $W = \exp(2\pi j/N)$. First, we note that for $\underset{\sim}{x}(t)$ real and Gaussian, the samples of $\underset{\sim}{y}_1$ will be complex and Gaussian. Next,

$$\begin{aligned} E\{\underset{\sim}{y}_1(f)\} &= \frac{1}{N}\sum_{t=0}^{N-1} E\{\underset{\sim}{x}(t)\}\, W^{-ft} \\ &= \begin{cases} \eta_x & f=0 \\ 0 & \text{otherwise} \end{cases} \end{aligned} \qquad (4)$$

Finally, the variance of the spectrum component at frequency f is equal to the corresponding term in the DFT of $R_{\underset{\sim}{x}}(\tau)$. For if $R_x(\tau) \xleftrightarrow{\mathfrak{F}} \alpha(f)$, then by (3)

$$\begin{aligned} E\{\underset{\sim}{y}_1^{*}(f_1)\underset{\sim}{x}(t)\} &= \frac{1}{N}\sum_{u=0}^{N-1} E\{x^{*}(u)\,x(f)\}\, W^{uf_1} \\ &= \frac{1}{N}\sum_{u=0}^{N-1} R_{\underset{\sim}{x}}(t-u)\, W^{uf_1} \\ &= \alpha(f_1)\, W^{f_1 t} \end{aligned}$$

Then

$$E\left\{\underset{\sim}{y}_1^*(f_1)\underset{\sim}{y}_1(f_2)\right\} = \frac{1}{N}\sum_{u=0}^{N-1} E\left\{\underset{\sim}{y}_1^*(f_1)\,x(u)\right\} W^{-uf_1}$$

$$= \alpha(f_1)\frac{1}{N}\sum_{u=0}^{N-1} W^{u(f_1-f_2)} \qquad (5)$$

$$= \begin{cases} \alpha(f_1) & f_1 = f_2 \\ 0 & f_1 \neq f_2 \end{cases}$$

Since the variance of $\underset{\sim}{y}_1(f)$ is $E\left\{\underset{\sim}{y}_1^*(f)\underset{\sim}{y}_1(f)\right\}$, this completes the proof. (These results are applicable to the DFT of any stochastic process and will be used again in Section c, below.) We note that the samples of $\underset{\sim}{y}_1$ are independent, since they are Gaussian and, by (5), orthogonal.

In our case, $\eta_{\underset{\sim}{x}} = 0$ so $\underset{\sim}{y}_1$ is zero mean. $R_{\underset{\sim}{x}}(\tau) = \sigma_{\underset{\sim}{x}}^2\,\delta(\tau)$; hence $\alpha(f) = \sigma_{\underset{\sim}{x}}^2$ (a constant). Thus $\underset{\sim}{y}_1(f)$ is stationary with variance $\sigma_{\underset{\sim}{x}}^2$. Because the samples of $\underset{\sim}{y}_1$ are independent, the autocorrelation function of $\underset{\sim}{y}_1$ is

$$R_{\underset{\sim}{y}_1}(\phi) = \sigma_{\underset{\sim}{x}}^2\,\delta(\phi) \qquad (6)$$

c. Absolute Value. We next consider $\underset{\sim}{y}_2$, the complex absolute value of $\underset{\sim}{y}_1$. Since the complex-absolute-value operation is zero memory, it follows that if the samples of $\underset{\sim}{y}_1$ are independent, then so are the samples of $\underset{\sim}{y}_2$. It is well known that the absolute value of a zero-mean complex Gaussian signal has a Rayleigh density:

$$f(Y_2) = \frac{Y_2}{\sigma_{\underset{\sim}{x}}}\exp\left(-Y_2^2/2\sigma_{\underset{\sim}{x}}^2\right)U(Y_2) \qquad (7)$$

Where U(v) is the unit step,

$$U(Y) = \begin{cases} 1, & Y \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

The mean and variance of $\underset{\sim}{y}_2$ are

$$\underset{\sim}{\eta}_{y_2} = \sigma_{\underset{\sim}{x}} \sqrt{\frac{\pi}{2}} \tag{8}$$

$$\sigma^2_{\underset{\sim}{y_2}} = (2 - \pi/2)\, \sigma^2_{\underset{\sim}{x}} \tag{9}$$

The autocorrelation function of y_2 is found as follows:

By definition, $$R_{\underset{\sim}{y_2}}(\phi) = E\{y_2(f)\; y_2(f-\phi)\} \tag{10}$$

For ϕ nonzero, the samples are independent; hence

$$\begin{aligned} R_{\underset{\sim}{y_2}}(\phi) &= E\{\underset{\sim}{y_2}(f)\}E\{\underset{\sim}{y_2}(f-\phi)\} \\ &= \eta_{\underset{\sim}{y_2}}^2 \end{aligned} \tag{11}$$

For $\phi = 0$, $$\begin{aligned} R_{\underset{\sim}{y_2}}(\phi) &= E\{[\underset{\sim}{y_2}(f)]^2\} \\ &= \sigma_{\underset{\sim}{y_2}}^2 + \eta_{\underset{\sim}{y_2}}^2 \end{aligned} \tag{12}$$

Hence $$R_{\underset{\sim}{y_2}}(\phi) = \eta_{\underset{\sim}{y_2}}^2 + \left(\sigma_{\underset{\sim}{y_2}}^2 + \eta_{\underset{\sim}{y_2}}^2\right)\delta(\phi) \tag{13a}$$

$$= \sigma_{\underset{\sim}{x}}^2\left[\frac{\pi}{2} + 2\,\delta(\phi)\right] \tag{13b}$$

d. Root Compression (y_3). In considering the n^{th}-root function, $\underset{\sim}{y}_3(f)$, the situation becomes more complicated. Since taking the n^{th} root is also a zero memory operation, the autocorrelation of $\underset{\sim}{y}_3$ follows the reasoning of equations 10 through 13. The values of $\eta_{\underset{\sim}{y}}$ and $\sigma_{\underset{\sim}{y}}$ are found from the density function of $\underset{\sim}{y}_3$. Following Papoulis (Ref. 2, p. 126f) (and using his notation), we have

$$y = g(x) = x^{1/n}$$

$$x_1 = g^{-1}(y) = y^n$$

$$g'(x) = \frac{1}{n} x^{\left(\frac{1-n}{n}\right)}$$

$$g'(x_1) = 1/ny^{n-1}$$

Sincc

$$f_{\underset{\sim}{y}}(y) = f_{\underset{\sim}{x}}(x_1) / \left| g'(x_1) \right| ,$$

and
$$f_{\underset{\sim}{x}}(x) = \frac{x}{\sigma_{\underset{\sim}{x}}^2} \exp\left(-x^2/2\sigma_{\underset{\sim}{x}}^2\right) U(x)$$

the required density function is

$$f_{\underset{\sim}{y}}(y) = \frac{n}{\sigma_{\underset{\sim}{x}}^2} y^{2n-1} \exp\left(-y^{2n}/2\sigma_{\underset{\sim}{x}}\right) U(y)$$

Returning to our own notation,

$$f_{\underset{\sim}{y}_3}(y_3) = \frac{n}{\sigma_{\underset{\sim}{x}}^2} y_3^{2n-1} \exp\left(-y_3^{2n}/2\sigma_{\underset{\sim}{x}}^2\right) U(y_3) \qquad (14)$$

We also need the moments of this density function.

The mth moment is given by

$$m_m = \int_{-\infty}^{\infty} y_3^m f_{\underset{\sim}{y}_3}(y_3)\, dy_3$$

$$= \frac{n}{\sigma_{\underset{\sim}{x}}^2} \int_{-\infty}^{\infty} y_3^{2n+m-1} \exp\left(-y_3^{2n}/2\sigma_{\underset{\sim}{x}}^2\right) dy_3$$

This can be evaluated by a change of vaiiable: with

$u = y_3^{2n}/2\sigma_{\underset{\sim}{x}}^2$,

$$m_m = \left(2\sigma_{\underset{\sim}{x}}^2\right)^{\frac{m}{2n}} \int_0^{\infty} u^{\frac{m}{2n}} e^{-u}\, du$$

Integrals oi this form are tabulated in Ref. 3. The result is

$$m_m = \left(2\sigma_{\underset{\sim}{x}}^2\right)^{\frac{m}{2n}} \Gamma\left(\frac{m+2n}{2n}\right) \qquad (15)$$

Fron these derivations we have t'e following particular results:

1. Area $A = m_0 = (2\sigma_{\underset{\sim}{x}}^2)^0\, \Gamma(1) = 1$ (not surprising)

2. Mean $\eta_{\underset{\sim}{y}_3} = m_1 = (2\sigma_{\underset{\sim}{x}}^2)^{\frac{1}{2n}}\, \Gamma\left(\frac{1+2n}{2n}\right)$

3. Second moment: $m_2 = (2\sigma_{\underset{\sim}{x}}^2)^{1/n}\,\Gamma\left(\frac{1+n}{n}\right)$

4. Variance: $\sigma_{\underset{\sim}{y}_3}^2 = (2\sigma_{\underset{\sim}{x}}^2)^{\frac{1}{n}}\left\{\Gamma\left(\frac{1+n}{n}\right) - \Gamma^2\left(\frac{1+2n}{2n}\right)\right\}$

5. $\lim\limits_{n\to\infty} \eta_{\underset{\sim}{y}_3} = 1$

6. $\lim\limits_{n\to\infty} \sigma_{\underset{\sim}{y}_2}^2 = 0$

7. $\eta_{\underset{\sim}{y}_3}^2/\sigma_{\underset{\sim}{y}_3}^2 = \dfrac{\Gamma^2\left(\frac{1+2n}{2n}\right)}{\Gamma\left(\frac{1+n}{n}\right) - \Gamma^2\left(\frac{1+2n}{2n}\right)}$

The density function for $\underset{\sim}{y}_3(f)$ is plotted for several different values of n in Figure 3.

Using these results and equation 10, we have, for the n^{th}-root function

$$R_{\underset{\sim}{y}_3}(\phi) = \eta_{\underset{\sim}{y}_3}^2 + \sigma_{\underset{\sim}{y}_3}^2\,\delta(\phi) \qquad (16a)$$

$$= (2\sigma_{\underset{\sim}{x}}^2)^{1/n}\left\{\Gamma\left(\frac{1+2n}{2n}\right) + \left[\Gamma\left(\frac{1+n}{n}\right) - \Gamma^2\left(\frac{1+2n}{2n}\right)\right]\delta(\phi)\right\} \qquad (16b)$$

In the limit as n increases indefinitely,

$$\lim_{n\to\infty} R_{\underset{\sim}{y}_3}(\phi) = 1$$

e. Second Transform $(\underset{\sim}{z})$. We finally turn to $\underset{\sim}{z}(v)$, the DFT of $\underset{\sim}{y}_3$. Using the results of Section b above, and particularly Equations (4) and (5), $\underset{\sim}{z}$ will again be a sequence of independent random variables such that $E\{\underset{\sim}{z}(v)\} = \eta_{\underset{\sim}{y}_3}\delta(v)$ and $E\{\underset{\sim}{z}^2(v)\} = \alpha(v)$ where α is the transform of (16):

$$\alpha(v) = \sigma^2_{\underset{\sim}{y}_3} + \eta_{y_3}^2\,\delta(v)$$

Hence

$$z(v) = \underset{\sim}{a}\,\delta(v) + \underset{\sim}{b} \qquad (17)$$

where $\underset{\sim}{a}$ and $\underset{\sim}{b}$ are random amplitudes with the statistics,

$$E\{\underset{\sim}{a}\} = \eta_{\underset{\sim}{y}_3}, \quad E\{\underset{\sim}{a}^2\} = \eta^2_{\underset{\sim}{y}_3}$$

$$E\{\underset{\sim}{b}\} = 0, \quad E\{\underset{\sim}{b}^2\} = \sigma^2_{\underset{\sim}{y}_3}$$

f. Removal of Simplifications. At this point, we can remove the two principal simplifications in our model. First, the top half of the spectrum is set to zero (Step 5, p. 3). This is equivalent to multiplying $\underset{\sim}{y}_3$ by a rectangular window, and the effect will be to convolve $\underset{\sim}{z}(v)$ with a sin v/v function. This will show up as a broadening of the impulse at the origin, but will have little other visible effect on $\underset{\sim}{z}(v)$. Second, $\underset{\sim}{x}(t)$ is subjected to triangular weighting. This will convolve $\underset{\sim}{y}_1(f)$ with $(\sin f/f)^2$ and the resultant broadening of the peaks in $\underset{\sim}{y}_2$ and $\underset{\sim}{y}_3$ will cause a high-end falloff in $\underset{\sim}{z}(v)$. (Strictly, z(v) is multiplied by the Fourier transform of $(\sin f/f)^{2/n}$; this transform cannot be expressed in closed form.)

2.3 Comparison with Observed Results

Figure 4 shows $\underset{\sim}{z}(v)$ as plotted from an actual run of INTEL with noise input. The similarity with the theoretical result is apparent. The most conspicuous feature is the buildup at the origin. In the light of our analysis, we know that this is the sin v/v shape resulting from the convolution just discussed. The first few sidelobes of sin v/v are in fact clearly visible at "A" in the figure. The rest of the plot shows low-level noise, which corresponds to the term $\underset{\sim}{b}$ in equation (17).

The noise-removal operation corresponds to removing the buildup at the origin. Naturally, the entire buildup cannot be removed, especially not the component at $v = 0$, because this corresponds to the constant portion of $\underset{\sim}{y}_3(f)$. If the constant were set to zero, $\underset{\sim}{y}_3$, would frequently go negative. Since $\underset{\sim}{y}_3$ was derived from an absolute value, this would be absurd. Hence only the side lobes of the buildup are removed. The principal result of this operation is that $\underset{\sim}{y}_3'$ (the prime indicates the value of the function on the "return trip" through the process) will have a mean which is less than 1. When $\underset{\sim}{y}_2'$ is computed, its mean will be still lower, since its mean is approximately the n^{th} power of the mean of $\underset{\sim}{y}_3'$. When the phase is restored to $\underset{\sim}{y}_2'$ to generate $\underset{\sim}{y}_1'$, the result will be complex noise samples of much smaller amplitude than the samples of $\underset{\sim}{y}_1$. Hence a reduction in noise results.

2.4 Analysis with Speech Input

It remains to ask why speech is not simultaneously affected by this process. The answer, briefly, is that speech (at least, vocalic speech) is a periodic function rich in harmonics. The result is that z contains many components away from the origin which are unaffected by the gating applied about the origin.

The spectrum of vocalic speech consists of a train of harmonics spaced at the pitch frequency f_p. If we transform a short segment of such speech, weighted by a time window function w(t), then the result will be

$$Y_1(f) = \sum_k a_k W(f-kf_p)$$

where f_p is the pitch frequency,

W(f) is the transform of the time window used,

a_k is the complex amplitude of the k^{th} harmonic.

W(f) can be assumed real without loss of generality, and under the processing conditions used, the overlap between harmonics is negligible. If there is no overlap, then at any frequency f the only contribution is that of the nearest harmonic. Because of this fact, we can take absolute values and n^{th} roots inside the summation: Hence

$$y_2(f) = |y_1(f)| \approx \sum_k |a_k| \, |w(f-kf_p)| \qquad (18)$$

and

$$y_3(f) = [y_2(f)]^{1/n} = \sum_k |a_k|^{1/n} W'(f-kf_p) \qquad (19)$$

where W'(f) is the n^{th} root of W(f). Hence after the n^{th}-root process, $y_3(f)$ is still periodic. This means that after the second transformation, z(v) will have at least one component not at the origin, corresponding to the period of $y_3(f)$. Indeed, z(v) has many components because y_3 is not a sinusoid. Thus

$$z(v) = \sum_m w_m A(v - mv_p) \qquad (20)$$

where the spacing $v_p = 1/f_p$ and the amplitudes w_m depend on the shape of W'(f). Specifically, if the Fourier transform of W'(f) is w'(v), then $w_m = w'(mv_p)$. The w_m values will depend on the specific function W(f) and on how taking the n^{th} root of W affect the shape of w'(v). A(v) is the transform of the envelope of $y_3(f)$ and thus contains formant information and some talker-identity cues.

Although the n^{th}-root operation is nonlinear, and hence superposition does not apply, in practice the presence of noise does not prevent this periodic structure from showing up. When the part of z(v) around the origin is gated out, this also affects the amplitude of the speech signal. In general, however, the peak shape A(v), which contains the formant information, is significantly wider than sin v/v, so a fair amount of it escapes the gating process. (In fact, one of the considerations determining the width of the gate is that as much of A(v) shall be preserved as possible.) Between the surviving portion of A(v) and the components found about multiples of v_p, enough information is preserved to provide recognizable speech in the output signal.

A side effect of the gating process is that it produces a slight enhancement of the high end of $y_3'(f)$ (above 2.2 kHz) and a region of attenuation from about 1.2 kHz to 2 kHz. When y_3' is raised to the n^{th} power to form y_2', these effects are exaggerated and the quality of the recovered speech is degraded. To counteract this effect, we multiply y_2' by an equalizing function. (It might be possible also to correct this distortion by shaping the edges of the gating function but as post-equalization works well enough, we have not tried doing so.)

2.5 Logarithmic Compression

A question which arises in the course of this analysis is why compression by means of the root compression works while logarithmic compression does not. Using the background provided by the foregoing analysis, it is easy to show that a logarithmic transformation does not produce the separation of the noise components that the n^{th} root does.

If a logarithmic conversion is substituted for the root operation, the statistics of $\underset{\sim}{y}_3(f)$ are different. If the log conversion is given by

$$y_3 = k \ln y_2,$$

then the probability density function of $\underset{\sim}{y}_3$ is

$$f_{\underset{\sim}{y}_3}(y_3) = \frac{1}{k\sigma_{\underset{\sim}{x}}^2} \exp\left(\frac{2y_3}{k}\right) \exp\left[\frac{-1}{2\sigma_{\underset{\sim}{x}}^2} \exp\left(\frac{2y_3}{k}\right)\right] \tag{21}$$

(This is determined the same way as before.) This function is plotted for various values of k with $\sigma_{\underset{\sim}{x}} = 1$ in Figure 5. The mth moment of this density is

$$M_m = \frac{1}{k\sigma_{\underset{\sim}{x}}^2} \int_{-\infty}^{\infty} y_3^m \exp\left(\frac{2y_3}{k}\right) \exp\left[\frac{-1}{2\sigma_{\underset{\sim}{x}}^2} \exp\left(\frac{2y_3}{k}\right)\right] dy_3.$$

By a suitable change of variable, this integral can be evaluated for m = 1; it does not converge for higher m. The mean is given by

$$\eta_{\underset{\sim}{y}_3} = k\left[\ln(2\sigma_{\underset{\sim}{x}}^2) - C\right] \tag{22}$$

when C is Euler's constant ≈.5772156649. The mean is a linear function of k and therefore vanishes as k approaches zero. The st ndard deviation, however, is not finite.

Notice that here we have a situation that is the opposite of the INTEL case. For INTEL,

$$\underset{\sim}{z} = \underset{\sim}{a}\ \delta(v) + \underset{\sim}{b}$$

where for increasing n, $E\left\{a^2\right\} \to 1$ and $E\left\{b^2\right\} \to 0$. Here, however, $E\left\{a^2\right\} \to 0$ and $E\left\{b^2\right\} \to \infty$; hence we get noise everywhere and no buildup at the origin. Since we rely on this buildup to make the noise components separate, log conversion does not lead to the desired result.

In actual computation, of course, it is not practical to do a pure log conversion since zero spectrum samples can occur. Instead, below some selected value a straight line approximation is used. This probably prevents the second moment of $f_{\underset{\sim}{y}_3}(y_3)$ from blowing up as described above, but it does not alter the fact that $\sigma_{\underset{\sim}{y}_3}{}^2$ is large compared to $n_{\underset{\sim}{y}_3}{}^2$ and that therefore no buildup occurs.

2.6 Optimization

Although these studies provide an understanding of the processes involved, they do not yield tidy, quantitative answers to such questions as how best to choose the root compression factor and how much improvement this optimum will yield. A little reflection will show that the answer, if available, would be of little use. We know empirically that factors less than 2 or greater than 4 are unsatisfactory. The evidence also indicates that the variation in quality, as n is varied over the region from 2 to 4, is not great. It seems highly unlikely, in view of the distribution functions uncovered in the analysis, that there is some sharp (i.e., narrow) optimum hidden somewhere in this range. The analysis suggests, in fact, that other lines of attack may prove more fruitful. These will be discussed in Section 4.

3.0 FURTHER DEVELOPMENT OF THE INTEL PROCESS

Under the theoretical studies, described in Section 2, we examined the way root-compression followed by Fourier transformation of the spectrum improves the separability of speech and additive noise. During experimental studies, described in this section, we tried to find additional ways to attenuate the components of noise, either in the spectrum or in the second-order spectrum, without at the same time equally attenuating the components of speech. Unfortunately, we did not succeed. Nevertheless, the techniques we examined are worth describing since they illustrate the kinds of processing that have been tested in our search for methods to improve the INTEL process.

3.1 Threshold Clipping

As discussed earlier, the INTEL process enhances the S/N of a signal in the second-order spectrum by attenuating a region in which the S/N is lower than the overall value. Alternatively, the S/N can be raised by emphasizing regions of the second-order spectrum in which the S/N was higher than the overall value. Such regions occur at integral multiples of the period of the vocal pitch, at which locations speech power concentrates in the second-order spectrum. Based on this approach, we developed and tested a method of processing the transformed signal, which we named Pitch Zone Emphasis, and which is described in a previous report.* As implied by the name, the method was to emphasize components of the second-order spectrum in the neighborhood of multiples of the pitch period. It proved to be very effective when the pitch was known with a maximum error of ±10 percent. However, when the error exceeded this limit, the process emphasized components of noise rather than those of speech, thereby making the output worse than the input. Obviously, this approach to processing the transformed signal, while conceptually correct, cannot be used until improved methods of measuring pitch at S/N below 0 dB are available.

Under the current study, we explored another method of emphasizing the speech components in the second-order spectrum. This new method does not attempt to locate the regions in which speech components are concentrated. Instead it takes advantage of the observation that the amplitudes of these components tend to be larger than the amplitudes of noise in the same regions.

Using this approach, we determine the average level of the noise at all points in the second-order spectrum. This

* Final report on Contract F30602-73-C-0100

average-level function is used as a threshold for discriminating between speech and non-speech components. We can use this threshold in either of two ways to emphasize the speech components:

1. Samples smaller than the threshold are set to zero. Samples larger than it are unaffected.
2. The threshold is subtracted from the absolute amplitude of each sample in the psuedo cepstrum. (Samples smaller than the threshold are set to zero.)

For convenience, we refer to the function as a clipping threshold, the first method of using it as absolute clipping, and the second method as center clipping.

A series of tests were run for each of these approaches. The amplitude of the clipping threshold, which was held constant during each test, was varied from 0.2 to 2.0 times the average noise amplitude function.

The signal regenerated after absolute clipping of the second-order spectrum was almost indistinguishable from the signal regenerated without the use of absolute clipping. The only significant difference occurred for thresholds that were greater than 1.5 times the average noise amplitude function.

In these particular tests the acoustic quality of the noise was transformed from a steady hiss to a partial gurgling sound.

Center clipping did enhance the signal-to-noise ratio in the output signal. However, it also tended to flatten the envelope of the signal spectrum, thereby suppressing the ratio of peak-to-valley amplitudes of the formants. The effect was to make the regenerated speech distinctly less intelligible. This distortion is caused by the tendency of center clipping to emphasize the peaks that are centered at integral multiples of the pitch period. Pitch zone shaping avoided this form of distortion by giving equal emphasis to all components in a narrow range around this central peak.

The approach described above points toward a third possible method of using a clipping threshold. In this proposed method, the clipping threshold would be used to detect signal components that exceeded it. Presumably, if the clipping threshold was set properly, most of the time, these components would be those of speech. Whenever such a component was found, it, and all components within a narrow range around it, would be left unaltered. All other components would be attenuated. In this way the threshold would be used to detect potential pitch zones. However, no attempt would be made to identify the correct pitch period.

3.2 Pitch Harmonic Emphasis

One objective of our study was to examine an approach that attenuates the noise between the pitch harmonics in the spectrum of noisy speech. Obviously, this approach requires knowing where the pitch harmonics are, which is the same as knowing the pitch frequency. In effect, this method is equivalent to passing the speech signal through a comb filter in which the comb spacing is equal to the pitch frequency.

At the outset, it is apparent that this method cannot succeed if there is significant error in the pitch frequency measurement. For example, if the measurement is in error by five percent, the estimated location of the tenth harmonic will be in error by 50 percent of the measured pitch frequency. In other words, the harmonic would be located halfway between its true position and that of one or another of the adjacent harmonics. Unfortunately, the error frequently exceeds five percent for signals in which the intelligibility or quality is low enough to require some form of processing.

Even if the pitch frequency were known with perfect accuracy, the method of comb filtering would not produce a useful enhancement of the speech quality. Such a filter can only be used to pass components at multiples of the pitch periods. It will pass such components whether they are of speech, noise, or speech-plus-noise. Thus, at frequencies where the pitch harmonics were negligible, the comb creates artificial harmonics composed of the noise components that pass through the comb. In tests of this method we found that the artificial harmonics generated a buzzing sound that tracks the pitch of the speech, and that is far more objectionable than is the steady hiss of the input noise.

There is a second, more basic reason why comb filtering is ineffective in enhancing speech quality. In a pair of experiments we demonstrated that the noise that occurs at the frequencies of the harmonics degrades speech quality far more than does the noise that occurs between the harmonics. For these experiments we generated a comb of noise, that is, a noise signal in which the noise was confined to uniformly spaced bands that were spaced one pitch interval from center to center. For the first test we added this noise to speech, after arranging the noise bands so that they coincided with pitch harmonics. For the second test, we offset the bands so that they fell between the harmonics. In both tests the average level of the noise was made equal to that of the speech. The results clearly showed that the intelligibility was far greater for the second test than for the first one. In fact, it was possible to still understand most of the speech in the second test after the noise had been raised to a level that reduced the intelligibility to zero in the first test. Since comb filtering cannot attenuate the noise that coincides with the harmonics, it is of little potential value in enhancing the intelligibility of noise-obscured speech.

3.3 Processing of Narrow-Band Speech

As originally developed, INTEL was designed to process speech in a band from DC to 3500 Hz. This range corresponds to the width of most telephone channels. However, it sometimes happens that the bandwidth of the received speech is less than half this range. When such a signal is processed by INTEL, the relative amplitudes of pitch harmonics in the regenerated spectrum can be severely altered. Our objective was to develop a method of processing narrow-band speech without requiring alterations in the software that implements the INTEL procedure.

Through experimentation, we found that the original amplitude relationships among the harmonics was maintained if, before rooting the spectrum, we replaced the noise components in the spectrum region above the cut-off frequency of the speech by a DC level. This method worked best when the DC level was set equal to the average level of the spectrum below the cut-off frequency. After regenerating the spectrum, the recovered DC level was set to zero. Thus, not only does this method minimize distortion in the amplitudes of the pitch harmonics for narrow-band speech, but, by virtue of setting the band above the cut-off frequency to zero, it increases the S/N in the regenerated speech signal.

3.4 Investigation of Phase Effects

One of the most widely held beliefs in speech research is that the ear is insensitive to phase. By this is meant that for sine wave signals the ear can detect changes in amplitude and frequency but not changes in phase. The accuracy of this belief is well established. Not as well established but almost as widely held is the belief that the ear is insensitive to changes in the relative phases in the sinusoidal components of a complex signal such as speech. As a simple proof of the general truth of this belief, consider the apparent invariance in voice quality as a listener, who receives only reverberant speech sounds, moves about a live room.

There is one condition for which the ear is sensitive to phase, and that is when the relative phases of the components are changing rapidly. One instance of such a signal is noise.

In band-limited white noise the phase angle of each spectral component varies in a random manner with uniform probability over the range 0 to 360 degrees. If the random variation in phase is removed, then the characteristic quality of the noise sound will be transformed from a hiss to a noisy buzz. But a complex spectrum analysis of the noise will still exhibit a uniform distribution of components whose amplitudes vary in a Rayleigh manner.

Of greater relevance to our study is the fact that in the case of noisy speech the additive noise tends to randomize the phase angles of the speech components. The degree of randomization depends, of course, on the relative amplitudes of the noise and the speech component of interest. At the outset of the study we hypothesized that such a distortion of phase would cause a correlated distortion in speech quality that would contribute to the loss in speech intelligibility at low signal to noise ratios.

To test the hypothesis, we performed several experiments. In the first one we randomized the phase angles of the components of noise-free speech by substituting for them the phase angles of corresponding components of noise. The result was a distinct deterioration in the quality of the speech sounds which, although they were still fully intelligible, became harsh and unnatural.

We next performed an experiment that was the inverse of the first one. Noise was added to speech at a level equal to that of the speech. Then the phase angles of components in the original noise-free speech were substituted for those in the noisy speech. Comparison of the two signals showed clearly that the noisy speech with non-randomized phase angles was the more intelligible one. Subjectively it appeared to be louder against the background of added noise.

Finally, we used the procedure of the second experiment to restore non-random phase to the components of speech regenerated after INTEL processing of an input noisy speech signal. Here we substituted the phase angles of the noise-free speech for those in the complex spectrum of the processed signal before regeneration of the speech time waveform. When compared with the normal output of the INTEL process this signal was shown to be subjectively louder and more natural, with a corresponding increase in speech intelligibility.

None of the foregoing should be interpreted to imply that phase conveys information in speech signals. That it does not was demonstrated in an experiment in which the phases of speech components were imposed on those of noise in a signal that contained only noise. Not only was this signal not even remotely speech-like, it was indistinguishable in quality from the original signal.

What is apparent from these results is that rapid, random variations in phase degrade the naturalness of speech sounds. For speech at high S/N this degradation has no significant effect of speech intelligibility, any more than does the lack of naturalness in speech that has passed through non-linear phase networks or speech generated by vocoders. However, at S/N of about 0 dB, or where word intelligibility is about 30 percent, randomization of speech phase clearly contributes to a reduction in intelligibility beyond that caused by the presence of noise components.

4.0 CONCLUSIONS AND RECOMMENDATIONS

The theoretical analysis has provided the basis for an understanding of INTEL without, however, providing a basis for specific quantitative estimates. Nevertheless, it has suggested areas for possible future research that may possibly be more fruitful than the present experiments.

The experiments which were performed on this project were mostly useful in a negative sense. That is, we now know several modifications that don't do any good. These were experiments that had to be made sooner or later, however, so even though unsuccessful, they do not represent wasted effort.

The most promising area for further research now appears to be refinements in the gating of the low-quefrency region of the pseudo cepstrum. In particular,

1. The DC term itself should be attacked. We know it is risky to remove this component; we do not know whether it cannot be attenuated.
2. Since we know the noise contribution is a sin v/v function, we ought to be able to take advantage of this fact. If we cannot subtract sin v/v directly, because of non-linear effects in the root compression process, we ought to investigate

analogous processes, such as subtracting some simple function of sin v/v.

Second, methods of reducing phase noise should also be investigated. These include the following:

1. Use of the complex spectrum and complex pseudo-cepstrum instead of the amplitude-only versions of these functions. The complex functions retain the phase data of the input signal. It is possible that by low-quefrency filtering of the complex spectrum the phase data can be "enhanced" in the same way as the amplitude data is by the current form of INTEL.
2. Averaging of the phase angles within each spectrum peak. For noise-free speech, the phase angles in the central region of a harmonic peak vary linearly with frequency, with the amount of variation proportional to the pitch glide and the order of the harmonic. The average phase angle within a peak will be the same as the angle at the center of the peak. By averaging the phase angles across a peak in the spectrum of noisy speech we should be able to improve the estimate of the phase angle at the peak center.

3. By substituting an artificial phase function for the randomized phase function of the input signal. Such a function would have to be compatible with the pitch and rate of change of pitch of the input signal. Otherwise, the rate of change of phase at a pitch harmonic would not be compatible with the frequency of the harmonic in the speech spectrum. Obviously, to make these functions compatible, it will be necessary to measure the pitch of the input speech signal. However, it may be possible to tolerate some incompatibility at the high order, weaker amplitude pitch harmonics. Consequently, it may not be necessary to measure pitch with an accuracy greater than that achievable by existing techniques.

REFERENCES

1. M. R. Weiss and E. Aschkenasy, Automatic Detection and Enhancement of Speech Signals, Nicolet Scientific Corporation. (No date)

2. A. Papoulis, Probability, Random Variables, and Stochastic Processes, New York, McGraw-Hill, 1965.

3. I. S. Gradshteyn and I. M. Ryzhik, Table of Integrals, Series, and Products, New York (Academic Press), 1965

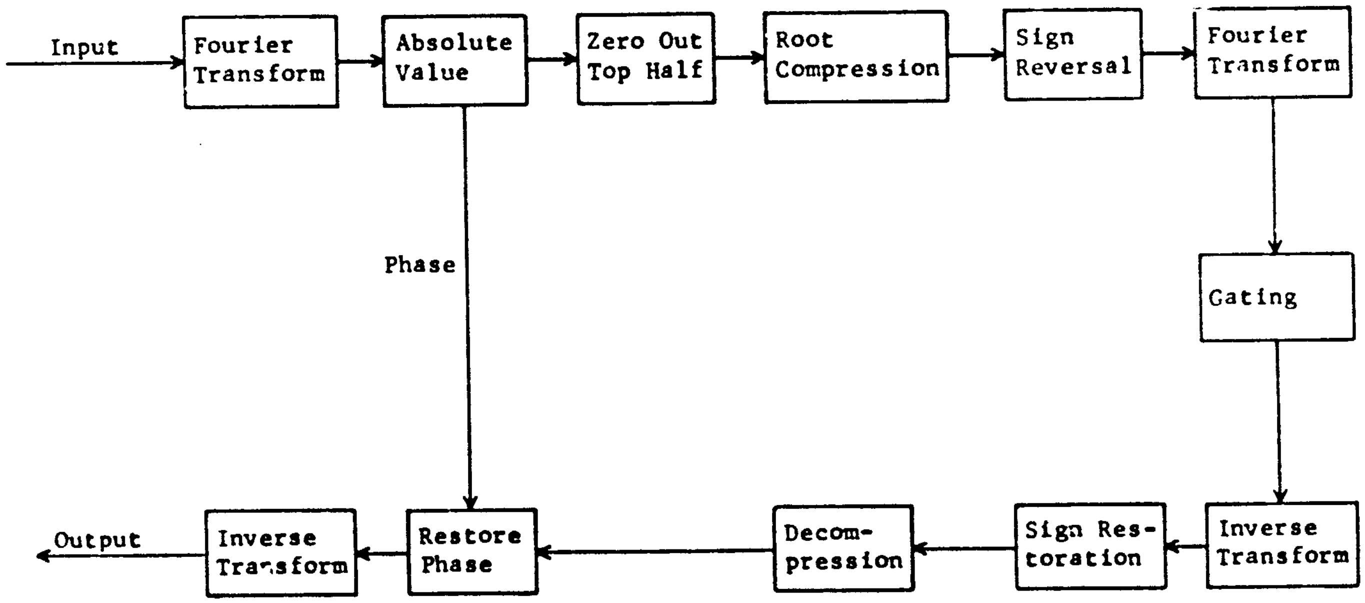

Figure 1. Block Diagram of INTEL

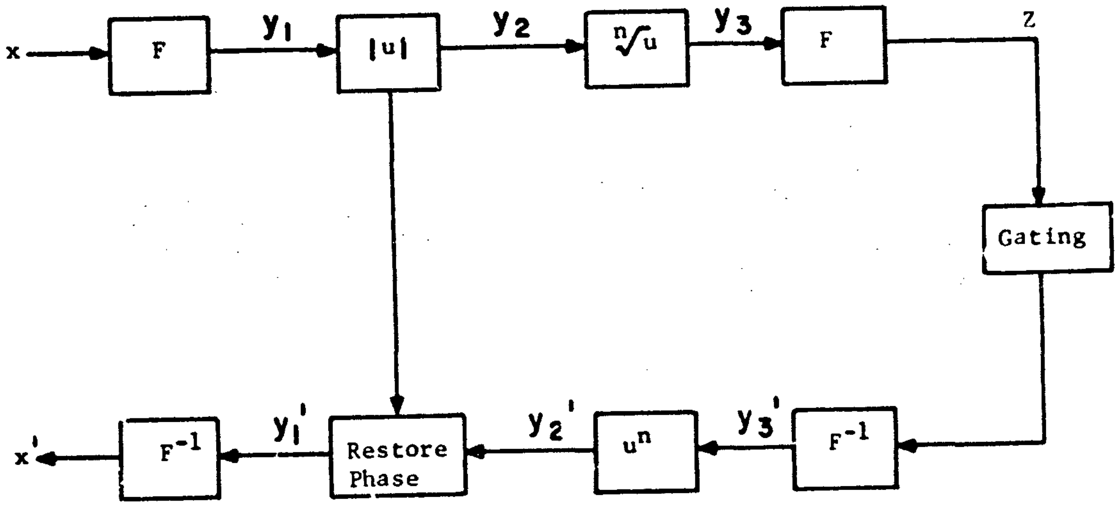

Figure 2. Simplified Model of INTEL

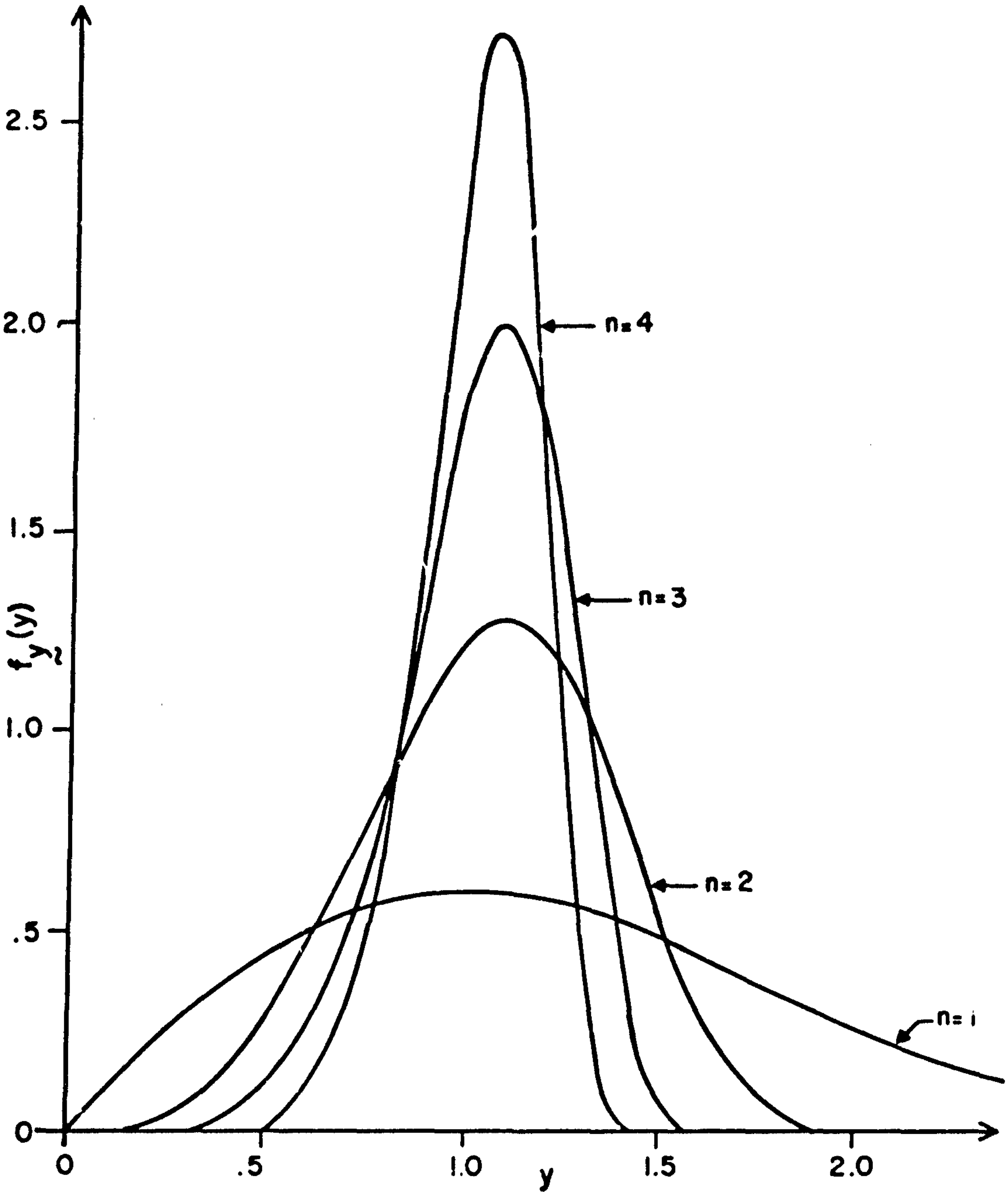

Figure 3. Probability Density Function of Root-Compressed Noise for Various Values of the Root-Compression Factor n

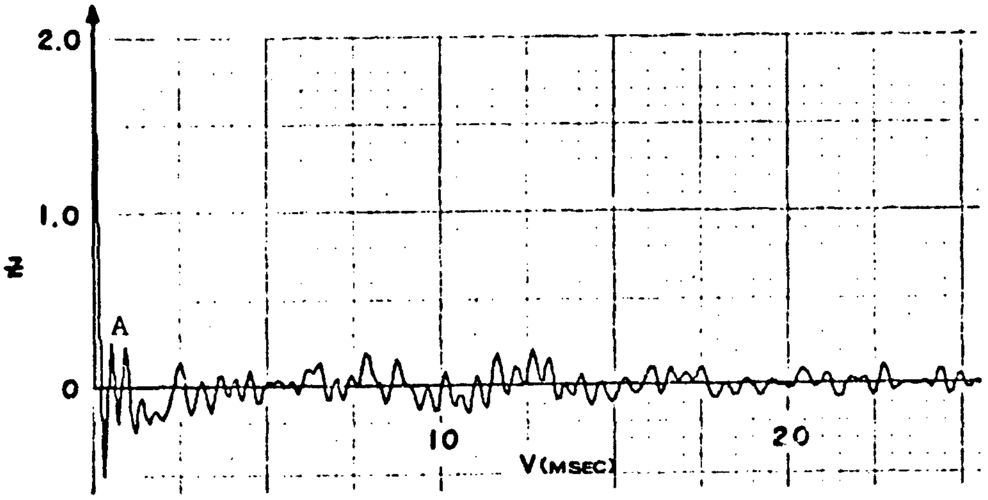

Figure 4. Appearance of Pseudo-Cepstrum

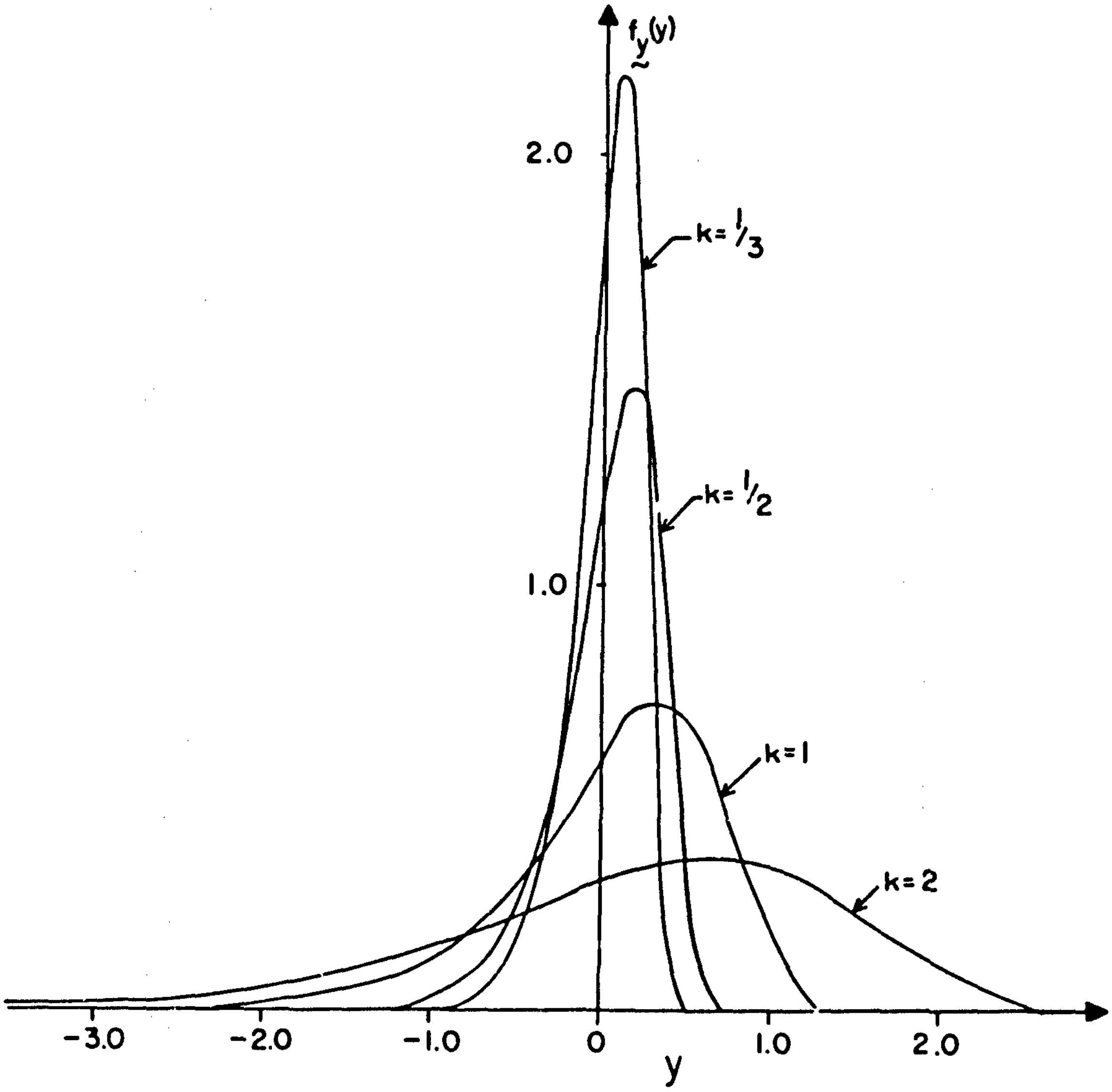

Figure 5. Probability Density Functior
of Logarithmically-Compressed Noise foı
Various Values of the Log-Compression Factor k

MISSION
of
Rome Air Development Center

RADC is the principal AFSC organization charged with planning and executing the USAF exploratory and advancec development programs for information sciences, intelligence, command, control and communications technology, products and services oriented to the needs of the USAF. Primary RADC mission areas are communications, electromagnetic guidance and control, surveillance of ground and aerospace objects, intelligence data collection and handling, information system technology, and electronic reliability, maintainability and compatibility. RADC has mission responsibility as assigned by AFSC for demonstration and acquisition of selected subsystems and systems in the intelligence, mapping, charting, command, control and communications areas.

END

DATE

FILMED

8-6-75

NTIS